

Protein databases

Henrik Nielsen

Protein databases, historical background

Swiss-Prot, <http://www.expasy.org/sprot/>

Established in 1986 in Switzerland

ExPASy (Expert Protein Analysis System)

Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

PIR, <http://pir.georgetown.edu/>

Established in 1984

National Biomedical Research Foundation, Georgetown University, USA

In 2002 merged into:

UniProt, <http://www.uniprot.org/>

A collaboration between SIB, EBI and Georgetown University.



UniProt

UniProt Knowledgebase (UniProtKB)

UniProt Reference Clusters (UniRef)

UniProt Archive (UniParc)

UniProt Knowledgebase Release 2012_05 (May 16, 2012) consists of:

UniProtKB/Swiss-Prot: Annotated manually (*curated*)

536,029 entries, 190,235,160 amino acids

UniProtKB/TrEMBL: Computer annotated (automatically translated from nucleotide databases)

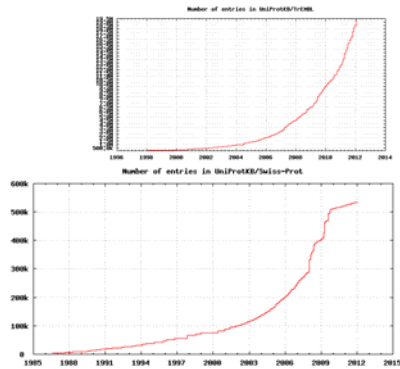
22,128,511 entries, 7,226,807,757 amino acids

Growth of UniProt

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENSCIENCE
LYSIS CBS

TrEMBL

Swiss-Prot



Content of UniProt Knowledgebase

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENSCIENCE
LYSIS CBS

- Amino acid sequences
- Functional and structural annotations
 - Function / activity
 - Secondary structure
 - Subcellular location
 - Mutations, phenotypes
 - Post-translational modifications
- Origin
 - organism: Species, subspecies; classification
 - tissue
- References
- Cross references

Amino acid sequences

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENSCIENCE
LYSIS CBS

From where do you get amino acid sequences?

- Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
- Direct amino acid sequencing: *Edman degradation*
- Mass spectrometry
- 3D-structures

UniProt entry, formatted view

CENTER FOR
RADIOLOGICAL
CALCULATIONS
ENCLAVE
LYSIS CBS

UniProt - Q5U0H1

Search in:

Download:

Accession # **Q5U0H1** **ELAD_HUMAN** **UniProt ID**

21-JUL-1986, integrated into UniProtKB/Swiss-Prot.

01-OCT-1996, sequence version 3.

25-JAN-2012, entry version 100.

DE RecName: Full=Alpha-1-antitrypsin;
DE AltName: Full=Alpha-1-protease inhibitor;
DE AltName: Full=Serpin A1;
DE Contains:
DE RecName: Full=Short peptide from AAT;
DE Short=SPAAAT;
DE Flags: Precursor;
GN Name=SERPINA1; Synonyms=AAT, PI; ORFNames=PRO0684, PRO2209;
OS Homo sapiens (Human);
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RZ REELINE=84107980; PubMed=6330997;
RA Bollen A., Herzig A., Cravedor A., Hertog P., Chuchana P.,
van der Straten A., Loris R., Jacobs P., van Elsen A.;
RT "Cloning and expression in Escherichia coli of full-length
complementary DNA coding for human alpha 1-antitrypsin."
RL DNA 2:255-264 (1983).

Protein names

Recommended name:
Alpha-1-antitrypsin
Alternative name(s):
Alpha-1-protease inhibitor
Alpha-1-antiprotease
Serpin A1
Cleaved into the following chain:
1 Short peptide from AAT
Short name: serpin

Gene names

Name: **SERPINA1**
Synonyms: AAT, PI
ORF Names: PRO0684, PRO2209

Organism

Human sapiens (Human)

Taxonomic identifier

NCBI:9606

Taxonomic lineage

Eukarya; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Protein attributes

UniProt entry, text view (flat file)

CENTER FOR
RADIOLOGICAL
CALCULATIONS
ENCLAVE
LYSIS CBS

```
ID      A1AT HUMAN          Reviewed:         418 AA.
AC      P01009; A6FZ14; B2RQ08; Q0PVP5; Q13672; Q532B6; Q5U0H1; Q7H4R2;
AC      Q6GUL6; Q6GUL9; Q6GFP3; Q6R851; Q6P1P0; Q6UC65; Q6UCR5;
DT      21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT      01-OCT-1996, sequence version 3.
DT      25-JAN-2012, entry version 100.
DE      RecName: Full=Alpha-1-antitrypsin;
DE      AltName: Full=Alpha-1-protease inhibitor;
DE      AltName: Full=Serpin A1;
DE      Contains:
DE      RecName: Full=Short peptide from AAT;
DE      Short=SPAAAT;
DE      Flags: Precursor;
GN      Name=SERPINA1; Synonyms=AAT, PI; ORFNames=PRO0684, PRO2209;
OS      Homo sapiens (Human).
OC      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC      Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC      Catarrhini; Hominidae; Homo.
OX      NCBI_TaxID=9606;
RN      [1]
RP      NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RZ      REELINE=84107980; PubMed=6330997;
RA      Bollen A., Herzig A., Cravedor A., Hertog P., Chuchana P.,
RA      van der Straten A., Loris R., Jacobs P., van Elsen A.;
RT      "Cloning and expression in Escherichia coli of full-length
RT      complementary DNA coding for human alpha 1-antitrypsin.";
RL      DNA 2:255-264 (1983).
...

```

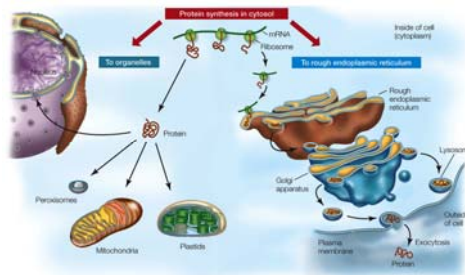
General annotation (Comments)

CENTER FOR
RADIOLOGICAL
CALCULATIONS
ENCLAVE
LYSIS CBS

General annotation (Comments)	
Function	Inhibitor of some proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The abundant form inhibits oxidant-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. (UniProt) (UniProt) (UniProt)
Subcellular location	Secreted (UniProt) Short peptide from AAT Secreted - extracellular space - extracellular matrix (UniProt)
Tissue specificity	Plasma
Domain	The reactive center loop (RCL) extends out from the body of the protein and directs binding to the target protease. The protease cleaves the serpin at the reactive site within the RCL, establishing a covalent linkage between the carboxyl group of the serpin reactive site and the amino hydroxyl of the protease. The resulting inactive serpin-protease complex is highly stable.
Post-translational modification	Several isomers are observed, resulting from the combination of different N-linked glycan structures and mature N-glycosylation. N-linked glycan at Asn-107 is alternatively 6-antennary, tri-antennary or tetra-antennary, whereas glycan at Asn-70 is 6-antennary with trace amounts of tri-antennary, and glycan at Asn-271 is exclusively 6-antennary. The structure of the antennae is Hex5Ac(alpha4-6)GlcNAc(beta4-4)GlcNAc attached to the core structure Man(alpha4-6)GlcNAc(beta4-4)GlcNAc. Some antennae are fucosylated, which forms a Lewis X determinant. Proteolytic processing may yield the truncated form that originates from Asn-30 to Lys-418.
Polymorphism	The sequence shown is that of the MTV allele which is the most common form of PI (M1 (44 to 49%). Other frequent alleles are: M1A (20 to 23%), M2 (10 to 11%), M3 (14 to 19%).
Involvement in disease	Defects in SERPINA1 are the cause of alpha-1-antitrypsin deficiency (A1ATD) (MIM:613400). A disorder whose most common manifestation is emphysema, which becomes evident by the third to fourth decade. A less common manifestation of the deficiency is liver disease, which occurs in children and adults, and may result in cirrhosis and liver failure. Environmental factors , particularly cigarette smoking, greatly increase the risk of emphysema at an earlier age. (UniProt) (UniProt) (UniProt)
Miscellaneous	The abundant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased.
Sequence similarities	Belongs to the serpin family.
Sequence caution	The sequence CACR0358.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened. The sequence CACR0356.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.

Protein sorting in eukaryotes

CENTER FOR
RADIOLOGICAL
CALCULUS
LYSIS CBS



Different proteins belong to different compartments of the cell – and some belong *outside* the cell

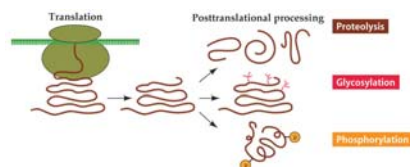
General annotation (Comments)

CENTER FOR
RADIOLOGICAL
CALCULUS
LYSIS CBS

General annotation (Comments)	
Function:	Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate effect for plasmin and thrombin. Inversely inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. (UniProt) (UniProt) (UniProt) Short peptide from AAT (SPAAT) is a reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE). (UniProt) (UniProt) (UniProt)
Subcellular location:	Secreted (UniProt) Short peptide from AAT: Secreted + extracellular space + extracellular matrix (UniProt)
Tissue specificity:	Plasma
Domain:	The reactive center loop (RCL) extends out from the body of the protein and directs binding to the target protease. The protease cleaves the serpin at the reactive site within the RCL, establishing a covalent linkage between the carboxyl group of the serpin reactive site and the serine hydroxyl of the protease. The resulting inactive serpin-protease complex is highly stable.
Post-translational modification:	Several isoforms are observed, resulting from the combination of different N-linked glycan structures and mature N-terminus. N-linked glycan at Asn-107 is alternatively di-antennary, tri-antennary or tetra-antennary, whereas glycan at Asn-70 is di-antennary with three amounts of bi-antennary, and glycan at Asn-271 is exclusively di-antennary. The structure of the antennae is Hex5A(GlcNAc)6(GlcNAc)4(GlcNAc) attached to the core structure Man(alpha1-6)Man(alpha1-3)Man(beta1-4)GlcNAc(beta1-4)GlcNAc. Some antennae are fucosylated, which forms a Lewis-X determinant.
Polymorphism:	Proteolytic processing may yield the truncated form that ranges from Asp-30 to Lys-410. The sequence shown is that of the MTV allele which is the most common form of P (44 to 45%). Other frequent alleles are: M1A 20 to 23%, M2 10 to 11%, M3 14 to 19%.
Involvement in disease:	Defects in SERPINA1 are the cause of alpha-1-antitrypsin deficiency (A1ATD) (MIM:613400). A disorder whose most common manifestation is emphysema, which becomes evident by the third to fourth decade. A less common manifestation of the deficiency is liver disease, which occurs in children and adults, and may result in cirrhosis and liver failure. Environmental factors, particularly cigarette smoking, greatly increase the risk of emphysema at an earlier age. (UniProt) (UniProt) (UniProt)
Macellaneous:	The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased.
Sequence similarities:	Belongs to the serpin family
Sequence caution:	The sequence CADC234.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened. The sequence CADC235.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.

Post-translational modifications

CENTER FOR
RADIOLOGICAL
CALCULUS
LYSIS CBS



Many proteins need to be *modified* after their synthesis to become active

Proteolysis: cleavage of *signal peptides*, *propeptides* or *initiator methionine*

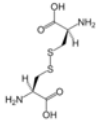
Glycosylation: Especially relevant on the cell surface. Also plays a role in sorting of proteins to *lysosomes*

Phosphorylation: Often *reversible*. Regulates the *activity* of many enzymes

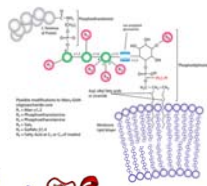
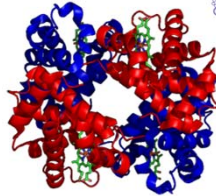
More post-translational modifications

CENTER FOR
BIOLOGICAL
CALCULUS
ENGINEERING
LYSIS CBS

- Lipid anchors
 - (e.g. GPI anchors)
- Disulfide bonds



- Prosthetic groups
 - (e.g. metal ions)



General annotation (Ontologies)

CENTER FOR
BIOLOGICAL
CALCULUS
ENGINEERING
LYSIS CBS

Ontologies

Keywords	
Biological process	Acute phase Blood coagulation Extracellular matrix Secreted
Cellular component	Extracellular space Polysaccharide
Coding sequence diversity	Signal
Domain	Protease inhibitor Serine protease inhibitor
Molecular function	Glycoprotein EC structure Complete proteome Secret protein sequencing Inference proteome
PTM	
Tactical term	
Gene Ontology (GO)	
Biological process	acute phase response protein activation protein degradation regulation of proteolysis extracellular space protein alpha-granule lumen extracellular matrix protein binding serine-type endopeptidase inhibitor activity
Cellular component	extracellular space protein alpha-granule lumen extracellular matrix
Molecular function	protein binding serine-type endopeptidase inhibitor activity



QuickGO - <http://www.ebi.ac.uk/QuickGO>

Sequence annotation (Feature Table)

CENTER FOR
BIOLOGICAL
CALCULUS
ENGINEERING
LYSIS CBS

Sequence annotation (Features)

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing					
<input type="checkbox"/> Signal peptide	1-24	24	CDS (1-24) CDS (25-418)		PRO_000002277
<input type="checkbox"/> Chain	25-418	394	Alpha 1-antitrypsin (CDS)		PRO_000002403
<input type="checkbox"/> Peptide	375-418	44	Short peptide from AAT		
Regions					
<input type="checkbox"/> Region	368-382	15	RCL		
Sites					
<input type="checkbox"/> Site	362-363	2	Reactive bond		
Amino acid modifications					
<input type="checkbox"/> Modified residue	256	1	S-cysteine/cysteine		
<input type="checkbox"/> Glycosylation	79	1	N-linked (GlcNAc...) (branching) CDS (1-24) CDS (25-418)		
<input type="checkbox"/> Glycosylation	102	1	N-linked (GlcNAc...) (branching) CDS (1-24) CDS (25-418)		
<input type="checkbox"/> Glycosylation	271	1	N-linked (GlcNAc...) (branching) CDS (1-24) CDS (25-418)		
Notes of variations					
<input type="checkbox"/> Alternative sequence	307-418	112	Missing in isoform 3		VSP_000000
<input type="checkbox"/> Alternative sequence	356-418	63	Amino acid MUTQ → VSP in isoform 2		VSP_000000
<input type="checkbox"/> Natural variant	4	1	S → L in Z-Mutagen CDS (1-24)		VAP_000070
<input type="checkbox"/> Natural variant	26	1	D → A in V-Mutagen CDS (25-418)		VAP_000079
<input type="checkbox"/> Natural variant	37	1	A → T in M-Mutagen CDS (25-418)		VAP_001038
<input type="checkbox"/> Natural variant	69	1	A → T in M-Mutagen CDS (25-418)		VAP_000080
<input type="checkbox"/> Natural variant	63	1	R → C in CDS (25-418) CDS (25-418)		VAP_000081

Secondary structure (Feature Table)

CENTER FOR
RADIOLOGICAL
CALCULUS
ENIGMA
LYSIS CBS

Secondary structure

Feature	Position	Count
Turn	48-50	3
Helix	61-69	10
Beta strand	70-72	3
Beta strand	74-76	3
Helix	78-89	12
Helix	94-103	10
Turn	108-110	3
Helix	113-127	15
Beta strand	136-145	11
Helix	152-160	9
Beta strand	166-169	5
Helix	174-188	15
Turn	189-191	3
Beta strand	206-215	10
Beta strand	218-230	3
Helix	234-236	3
Beta strand	239-237	10
Beta strand	239-256	18
Turn	257-260	4
Beta strand	261-268	8
Turn	269-271	3
Beta strand	272-279	8

Protein structure

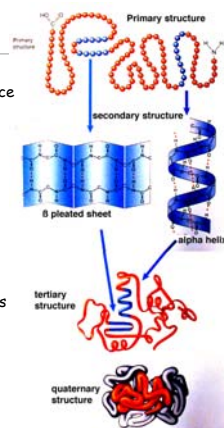
CENTER FOR
RADIOLOGICAL
CALCULUS
ENIGMA
LYSIS CBS

Primary structure: Amino acid sequence

Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

Quaternary structure: subunits



Evidence (Comments, Feature Table)

CENTER FOR
RADIOLOGICAL
CALCULUS
ENIGMA
LYSIS CBS

Q43495 (108_SOLLIC) Reviewed, UniProtKB/Swiss-Prot
Last modified March 2, 2010; Version 45; History...

General annotation (Comments)

Subcellular location: Secreted, **Extracellular**
Tissue specificity: Stamen- and tapetum-specific.
Sequence similarities: Belongs to the ADP/ATP family.

Sequence annotation (Features)

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing					
Signal peptide	1-30	30	Extracellular		
Chain	31-102	72	Protein 108		PRO_0000000238
Amino acid modifications					
Disulfide bond	41 ↔ 77		Extracellular		
Disulfide bond	51 ↔ 66		Extracellular		
Disulfide bond	67 ↔ 92		Extracellular		
Disulfide bond	79 ↔ 99		Extracellular		

CENTER FOR
RBIOLGICAL
CALSEQUENCE
ANALYSIS **CBS**

Sequence annotation and General comment:

-
-
-
-
-
-

CENTER FOR
RADIOBIOLOGICAL
SEQUENCE ANALYSIS **CBS**

CENTER FOR
RADIOBIOLOGY
CALSEQUENCE
ANALYSIS **CBS**

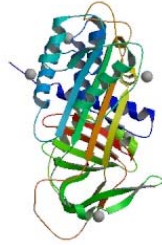
Cross-references, 3D structure

CENTERO
RIBILOGI
CALISEU
ENCEANA
LYSIS CBS

3D structure databases

- PDB
- RCSB PDB
- PDB

Entry	Method	Resolution (Å)	Chain	Proteins	PDBsum
1ATU	X-ray	2.70	A	45-418	[x]
1055	X-ray	3.00	A	44-377	[x]
1EDX	X-ray	2.60	B	378-418	[x]
			B	40-362	[x]
			B	303-418	[x]
1UD	X-ray	2.10	A	25-418	[x]
1WCT	X-ray	3.46	A	25-418	[x]
1OGB	X-ray	2.66	A	25-418	[x]
1QPH	X-ray	2.30	A	25-418	[x]
1PSI	X-ray	2.92	A	25-418	[x]
1QLP	X-ray	2.00	A	25-418	[x]
1QMB	X-ray	2.60	A	49-376	[x]
			B	377-418	[x]
2006	X-ray	3.30	A	26-362	[x]
			B	303-418	[x]
2000	X-ray	2.00	A	25-418	[x]
3CWC	X-ray	2.44	A	25-418	[x]
3CWM	X-ray	2.51	A	25-418	[x]
3CWM	X-ray	2.20	A	25-418	[x]
3C8U	X-ray	3.20	A/B/C	25-418	[x]
3NDD	X-ray	1.50	B	45-372	[x]
			B	303-418	[x]
3NDF	X-ray	2.70	B	45-361	[x]
			B	303-418	[x]
3T1P	X-ray	3.90	A	45-418	[x]
7AP1	X-ray	3.00	A	36-362	[x]
			B	303-418	[x]
8AP1	X-ray	3.10	A	36-362	[x]
9AP1	X-ray	3.00	B	303-418	[x]
			A	36-362	[x]
			B	303-418	[x]



Cross-references

CENTERO
RIBILOGI
CALISEU
ENCEANA
LYSIS CBS

Other databases linked from UniProt

(there are ~100 in total):

- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gene expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases

Translation and Reading Frames

CENTERO
RIBILOGI
CALISEU
ENCEANA
LYSIS CBS

The genetic code

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENSCIENCE
LYSIS CBS

		Second letter					
		U	C	A	G		
First letter	U	UUU Phenylalanine UUC Leucine UUA Leucine UUG Leucine	UCU Serine UCC Serine UCA Serine UCG Serine	UAU Tyrosine UAC Tyrosine UAA Stop codon UAG Stop codon	UGU Cysteine UGC Cysteine UGA Stop codon UGG Tryptophan	Third letter	U C A G
	C	CUU Leucine CUC Leucine CUA Leucine CUG Leucine	CCU Proline CCC Proline CCA Proline CCG Proline	CAU Histidine CAC Histidine CAA Glutamine CAG Glutamine	CGU Arginine CGC Arginine CGA Arginine CGG Arginine		U C A G
	A	AUU Isoleucine AUC Isoleucine AUA Isoleucine AUG Methionine start codon	ACU Threonine ACC Threonine ACA Threonine ACG Threonine	AAU Asparagine AAC Asparagine AAA Lysine AAG Lysine	AGU Serine AGC Serine AGA Arginine AGG Arginine		U C A G
	G	GUU Valine GUC Valine GUA Valine GUG Valine	GCU Alanine GCC Alanine GCA Alanine GCG Alanine	GAU Aspartic acid GAC Aspartic acid GAA Glutamic acid GAG Glutamic acid	GGU Glycine GGC Glycine GGA Glycine GGG Glycine		U C A G

- Degenerate (*redundant*) but not ambiguous
- Almost universal (deviations found in mitochondria)

Reading Frames 1

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENSCIENCE
LYSIS CBS

A piece of an mRNA-strand:

5' augccc aagcugaauagcguagagggguuuu ucaucauuugagga cgaugu auaa 3'

can be divided into triplets (*codons*) in three ways:

```

1 aug ccc aag cug aaU agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa
M P K L N S V E G F S S F E D D V *
2 ugc cca agc uga aua gcg uag agg ggu uuu cau cau uug agg acg aug uau
C P S * I A * R G F H H L R T M Y
3 gcc caa gcU gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua
A Q A E * R R G V F I I * G R C I

```

Each possible set of triplets is called a *reading frame*.

Reading Frames 2

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENSCIENCE
LYSIS CBS

Since there are two strands in DNA, there are *six* possible reading frames in a piece of DNA (three in each direction):

```

3 A Q A E * R R G V F I I * G R C I
2 C P S * I A * R G F H H L R T M Y
1 M P K L N S V E G F S S F E D D V *
5' ATGCCAAGCTGAATAGCTAGAGGGGTTTTCATCATTTGAGGACGATGTATAA 3'
3' TACGGGTTGACTTATCGCATCTCCCAAAAGTAGTAACTCTGCTACATATT 5'
H G L Q I A Y L P K * * K L V I Y L -1
G L S F L T S P N E D N S S S T Y -2
A W A S Y R L P T K M M Q P R H I -3

```

A reading frame from a start codon to the first stop codon is called an *open reading frame* (underlined above).